CRANBERRY: MEMORY-EFFECTIVE SEARCH IN 100M HIGH-DIMENSIONAL CLIP VECTORS

Vladimir Mic[1], Jan Sedmidubsky[2] and Pavel Zezula[2]

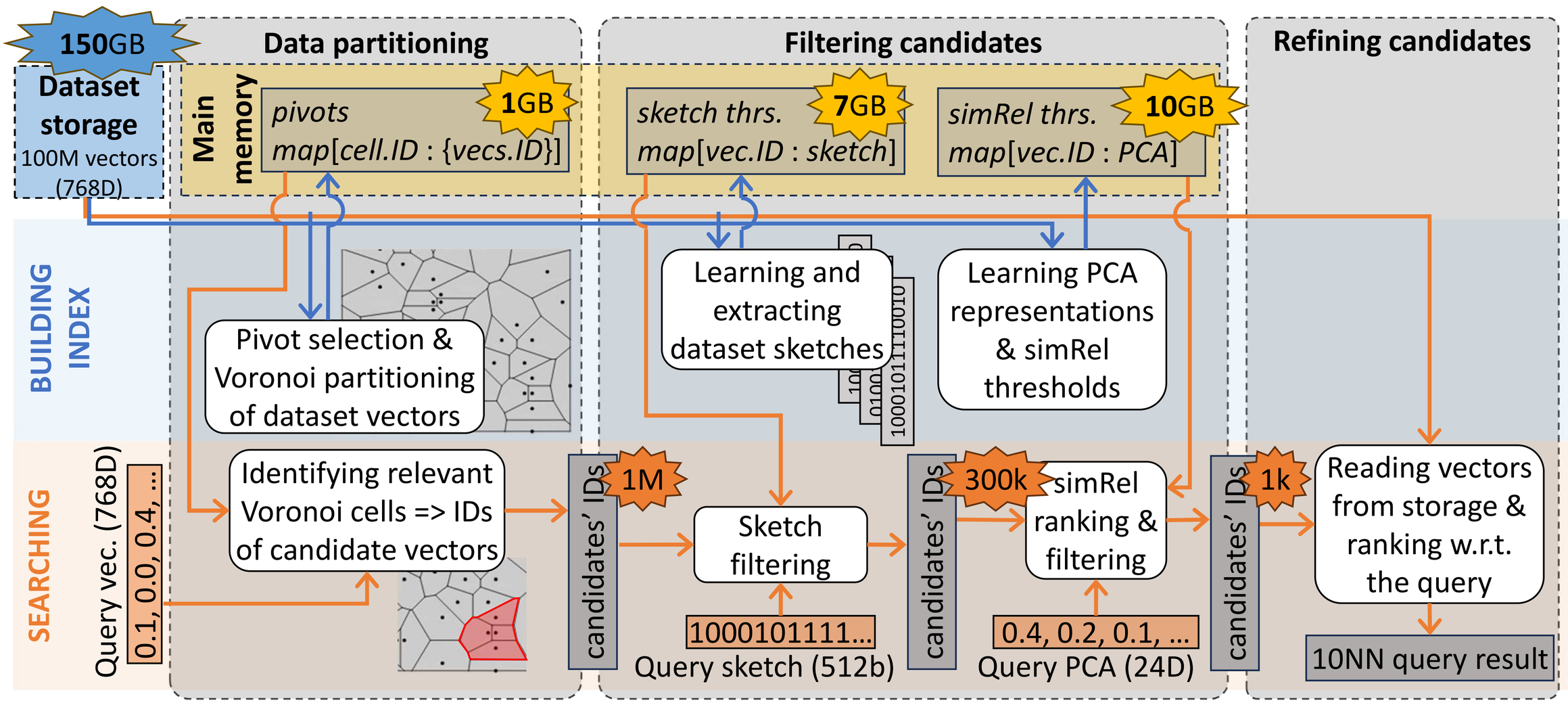
[1]Aarhus University, Denmark, [2]Masaryk University, Brno, Czech Republic

(1) OUR GOALS

- 10NN search in up to **100M CLIP Descriptors**
- Possibly stored **on a disk**
- **Refine** as **few vectors** as possible

(2) CRANBERRY OVERVIEW

- Data partitioning → the **Voronoi partitioning**
- Filtering of the partitions → **Binary Sketches + Relational Similarity + Early Termination**



(3) FILTERING POWER & ACCURACY (100M DATA)

	Voronoi P.		Voronoi P + Sketches		Voronoi P. + Sk. + simRel		all + early term. = CRANBERRY	
	Refined	Recall	Refined	Recall	Refined	Recall	Refined	Recall
Min	841,743	0	10	0	10	0	10	0
1st quart.	987,139	1.0	35,128	1.0	26,407	0.9	803	0.9
Median	993,611	1.0	192,890	1.0	75,631	1.0	836	1.0
3rd quart.	997,394	1.0	616,322	1.0	145,882	1.0	909	1.0
Max	1,000,000	1.0	1,000,000	1.0	627,375	1.0	1,390	1.0
Average	989,724	0.979	336,621	0.966	98,892	0.951	821	0.901

(4) SEARCH SPEED

Dataset size		100M	30M	10M
Refined		821	855	876
Recall		0.901	0.902	0.909
PC	Latency	1.07 s	0.40 s	0.16 s
96 GB RAM,	Throughput	4.7 q/s	12.6 q/s	30.9 q/s
data on SSD	$\sum$ time	2,139 s	796 s	324 s
Server	Latency	1.2 s	0.37 s	0.20 s
512 GB RAM,	Throughput	15.0 q/s	48.7 q/s	88.0 q/s
data in RAM	$\sum$ time	669 s	205 s	114 s

- The **Voronoi partitioning** with 20,000 pivots **discards 99.03 % of vectors**
  - and **preserves 97.9 %** of 10 true **nearest neighbours** per average query
- The **CRANBERRY refines** just **821 vectors** per average query (**0.0008 %** of the dataset), with the search accuracy of **90.1 %**
- **Bottlenecks for the SISAP Challenge**:
  - Dataset **fits into** 512 GB **main memory** – not our primary focus
  - **20,000 pivots** imply 20,000 distance computations to each query. **If the dataset is in RAM**, so many pivots are **unjustified**
  - The Voronoi partitioning identifies **1M candidates** which are **costly sorted by their Hamming dists.** to the query sketch
  - **Future work: replace the Voronoi partitioning** with some better data organisation